

A dependent partition-valued process for multitask clustering and time evolving network modelling

Konstantina Palla [†]
University of Cambridge

KP376@CAM.AC.UK

David A. Knowles [†]
University of Cambridge

DAVIDKNOWLES@CS.STANFORD.EDU

Zoubin Ghahramani
University of Cambridge

ZOUBIN@ENG.CAM.AC.UK

Abstract

The fundamental aim of clustering algorithms is to partition data points. We consider tasks where the discovered partition is allowed to vary with some covariate such as space or time. One approach would be to use fragmentation-coagulation processes, but these, being Markov processes, are restricted to linear or tree structured covariate spaces. We define a partition-valued process on an arbitrary covariate space using Gaussian processes and a novel interpretation of the stick breaking construction. By choosing the parameters of the stick breaking construction the process can be given Chinese restaurant process or Pitman Yor process marginals. We use the process to construct a multitask clustering model which partitions datapoints in a similar way across multiple data sources, and a time series model of network data which allows cluster assignments to vary over time. We use Elliptical Slice Sampling for inference and apply our method to defining cancer subtypes based on different types of cellular characteristics, finding regulatory modules from gene expression data from multiple human populations, and discovering time varying community structure in a social network.

1. Introduction

We are interested in problems where the partitioning of data into groups depends on some covariate. As a simple example, consider how the partitioning of a number of people into friendship groups might evolve over time (the covariate in this case). At proximal times people will tend to form similar partitions, but at more distant times the partitioning might be quite different. One of our examples in Section 9 will involve modelling such data. We define a nonparametric process that induces dependency between partitions on a covariate space.

Many nonparametric processes studied in the literature, such as the Dirichlet process (DP, [Ferguson, 1973](#)), are distributions over the space of measures. The DP is used to construct a distribution over the space of partitions known as the Chinese restaurant process (CRP, [Aldous, 1983](#)). Dependent nonparametric processes extend distributions over measures and partitions to give distributions over collections of measures or partitions indexed by locations in some covariate space ([MacEachern, 1999](#)). Covariate spaces include \mathbb{R}_+ (e.g. continuous time), \mathbb{Z} (e.g. discrete time), or \mathbb{R}^d (e.g. geographical location). Most of the dependent nonparametric processes explored in the literature define distributions over collections of measures. Dependency among the measures on each location may then be induced in various ways. The single-p Dependent Dirichlet Process (DDP, [MacEachern, 1999](#)) assumes a shared set of atom weights at each covariate index and induces dependency by allowing the corresponding atom locations to vary according to some stochastic process, e.g. a Gaussian process (GP, [Rasmussen & Williams, 2006](#)). If the covariate space is time, the process defines a nonparametric mixture

[†]These authors contributed equally.

model at each index and the characteristics of each component evolve over time. The multiple-p DDP by MacEachern (2000) allows the same atom locations at each covariate index but the weights of each atom are dependent across the indices. This is achieved by replacing the beta-distributed random variables in the stick-breaking construction with samples from a stochastic process evolving over the covariate space.

By de Finetti’s theorem, any exchangeable sequence is equivalent to i.i.d. draws from a conditional distribution and can be written as a mixture of such distributions. Both the single-p and multiple-p DDP have the DP as their de Finetti mixing measure and introduce dependency in the mixing distribution either through the atom locations or weights. The generalized spatial Dirichlet process by Duan et al. (2005) induces dependency by assuming that the mixing distribution is common to all covariate indices, but the conditional distributions are correlated. It defines a DP mixture model of Gaussian random fields at each covariate index.

Despite this long time of research into dependent measure-valued processes, little attention has been given to dependent *partition*-valued processes. A sample from such a process is collection of partitions indexed by the covariate: at any single covariate location, there is a single partition. An exception is Teh et al. (2011), where the duality between Kingman’s coalescent (KC, Kingman, 1982) and the Dirichlet diffusion tree (DDT, Neal, 2003a) is leveraged to define a “Fragmentation-Coagulation” process (FCP) which is Markov, stationary, exchangeable and has CRP distributed marginals (Bertoin, 2006). The FCP defines a distribution over a collection of partitions on a one dimensional covariate space. The partitioning at each covariate location is a result of fragmentation (according to the DDT) and coagulation (according to the KC) events that take place between adjacent covariate locations. Although mathematically elegant, due to its Markov construction it is not clear how to extend the FCP to an arbitrary covariate space. In this paper, we derive a dependent partition-valued process on an arbitrary covariate space which, like the FCP, is exchangeable and has CRP distributed marginals. For brevity, we refer this process as DPVP for “Dependent Partition-Valued Process”.

The DPVP is closely related to the dependent IBP (dIBP, Williamson et al., 2010), which addresses the problem of modelling dependence for binary latent feature models. Coupling over the covariate, in both processes, is achieved by representing Bernoulli variables at each covariate index as transformed Gaussian vari-

ables and aggregating these into Gaussian processes over the covariate. The dIBP generates a set of binary feature matrices evolving over the covariate, while the DPVP couples a set of CRPs.

We use the DPVP to construct two distinct models. The first is a multitask clustering model (MCM) which attempts to find similar partitions of objects across distinct data views. Our approach learns the similarity between the clustering in each data source. MCM is closely related to the Multiple Dataset Integration model (MDI, Kirk et al., 2012), where the conditional probability of allocating a sample to a cluster in one data source is influenced by the assignments in other data sources. Like MCM, MDI can learn how similar the clusterings should be across different data sources using positive real-valued parameters for each pair of data sources. However, MCM is a valid generative model unlike MDI which is defined only in terms of these conditional distributions. Both models can be considered as using finite approximations to a DP mixture model at each location (data source): where MDI simply uses a large finite Dirichlet distribution, MCM uses the stick breaking construction of the DP (Ishwaran & James, 2001b). Another distinction between the two approaches lies in the way the dependency across the partitions is induced. The MDI model introduces a positive real-valued parameter for each pair of data sources that describes their (dis)similarity, whereas MCM uses Gaussian processes to induce dependency between the partitions across the covariate space.

The outline of this paper is as follows. In Section 2, we briefly provide some background on partitions, the CRP and the stick breaking construction. In Section 3, we present the dependent partition-valued process and in Sections 4 and 5 describe how to use this process to build a multitask clustering model and a model for time evolving community structure respectively. In Section 6, we describe how different choices of kernel may be used for different applications. Inference in our model is performed via a Gibbs sampler, which is described in Section 7. In Section 8 and 9 we describe case studies for multitask clustering and network modelling. Finally, in Section 10 we conclude our work and discuss some future directions.

2. Background

A partition of $[N] = \{1, \dots, N\}$ is a set of disjoint non-empty subsets of $[N]$ such that the union of these subsets is $[N]$. In clustering applications we refer to each subset as a “cluster”. The set of partitions of $[N]$ is denoted Π_N . The most natural distribution over

Π_N is the Chinese restaurant process (CRP). Since the CRP is exchangeable (invariant to permutations of $[N]$) and projective (consistent under marginalisation of elements), by de Finetti's theorem it can be represented using i.i.d. samples from a random measure. For the CRP this is the Dirichlet process (DP). The weights of the DP can be represented using a "stick breaking" construction as follows [Sethuraman (1994); Ishwaran & James (2001a)]:

$$\begin{aligned}\pi_k &= v_k \prod_{l=1}^{k-1} (1 - v_l) & \forall k \in \mathbb{Z} \\ v_k &\sim_{\text{iid}} \text{Beta}(1, \alpha) & \forall k \in \mathbb{Z}\end{aligned}\quad (1)$$

In the above, the π_k 's are mixture proportions (stick lengths) and $\alpha > 0$ is known as the concentration parameter. If we now consider sampling cluster assignments $c_n | \pi \sim \text{Multinomial}(\pi)$ for $n \in [N]$, then the corresponding partition, γ of $[N]$ is marginally CRP distributed. The partition is obtained from the cluster assignments c by putting n and m in the same subset iff $c_n = c_m$.

3. A dependent partition-valued process

Given some covariate space \mathcal{T} our aim is to construct an exchangeable, projective, dependent process $(\gamma(t), t \in \mathcal{T})$ such that each $\gamma(t)$ is a random partition which is marginally CRP distributed. To achieve this we will use the following representation of the stick-breaking construction (Equation 1):

For each object $n \in [N]$

1. Set $k := 1$
2. Sample $a_{nk} \sim \text{Bernoulli}(v_k)$
3. If a_{nk} then assign $c_n := k$, else increment k and go to 2.

It is straightforward to see that the probability of choosing cluster k is given by the stick length π_k since

$$\begin{aligned}P(c_n = k | v) &= P(a_{n1} = 0 | v) \dots \\ &= P(a_{n(k-1)} = 0 | v) P(a_{nk} = 1 | v) \\ &= (1 - v_1) \dots (1 - v_{k-1}) v_k = \pi_k\end{aligned}\quad (2)$$

Following Williamson et al. (2010), we note that the Bernoulli random variable a_{nk} can be represented as

$$\begin{aligned}f_{nk} &\sim \mathcal{N}(0, \sigma^2) \\ a_{nk} &= \mathbb{I}[f_{nk} < \phi^{-1}(v_k | 0, \sigma^2)]\end{aligned}\quad (3)$$

where $\phi(\cdot | \mu, \sigma^2)$ is the Gaussian cumulative distribution function with mean, μ and variance, σ^2 . We now extend these random variables to random functions on \mathcal{T} and introduce dependency by extending the Gaussian prior on each f_{nk} to a Gaussian process (GP) prior on each $f_{nk}(t)$ with $t \in \mathcal{T}$:

$$\begin{aligned}f_{nk}(t) &\sim_{\text{iid}} \text{GP}(0, \Sigma(t, t')) & \forall n \in [N], k \in \mathbb{Z} \\ a_{nk}(t) &= \mathbb{I}[f_{nk}(t) < \phi^{-1}(v_k | 0, \Sigma(t, t))]\end{aligned}\quad (4)$$

where $\Sigma(t, t')$ is the covariance function which we assume to be common to all the GPs. As a result of the marginalisation properties of GPs, each $f_{nk}(t)$ is marginally $N(0, \Sigma(t, t))$ distributed, so that $a_{nk}(t)$ is marginally $\text{Bernoulli}(v_k)$ distributed, and the resulting partition $\gamma(t)$ is marginally CRP distributed, as desired.

To summarise, the DPVP generative process is

$$\begin{aligned}v_k &\sim_{\text{iid}} \text{Beta}(1, \alpha) & \forall k \in \mathbb{Z} \\ f_{nk} &\sim_{\text{iid}} \text{GP}(0, \Sigma) & \forall n \in [N], k \in \mathbb{Z} \\ c_n(t) &= \min_{f_{nk}(t) < \phi^{-1}(v_k | 0, \Sigma(t, t))} k\end{aligned}\quad (5)$$

We do not make the v_k a function on \mathcal{T} but instead assume global mixing proportions: relaxing this constraint would be a straightforward extension.

In practice, we cannot represent a countably infinite set of GPs. While we could adaptively extend our representation as required, we instead choose the simpler option of truncating the stick breaking construction at some level K . If $a_{nk}(t) = 0$ for all $k < K$, then we set $c_n(t) = K$. To sample from the K -truncated DPVP at T locations $\{t_\tau \in \mathcal{T} | \tau = 1, \dots, T\}$ we therefore require $K - 1$ T -vectors $\mathbf{f}_{nk} = [f_{nk}(t_1) \dots f_{nk}(t_T)]$, for each object n , from a Gaussian process with a $T \times T$ Gram (covariance) matrix, Σ , where $\Sigma_{\tau\tau'} = \Sigma(t_\tau, t_{\tau'})$. For notational simplicity, we concatenate the f vectors into a $T \times N(K - 1)$ matrix, \mathbf{F} . By drawing from a GP as in 3, we introduce dependency among the partitions in different covariate locations. The dependence is defined by the covariance function of the GP, Σ , and introduces similarity between the partitions at proximal covariate locations.

We denote a sample from the DPVP as $DPVP(\alpha, \mathbf{t}, \Sigma)$, where α is the concentration parameter of the underlying DP, \mathbf{t} is the vector of covariate locations and Σ is the Gram matrix. We now use this process to construct two models: a multitask clustering model and a network model that allows the discovered community structure to vary through time. The graphical model for both is shown in Figure 1.

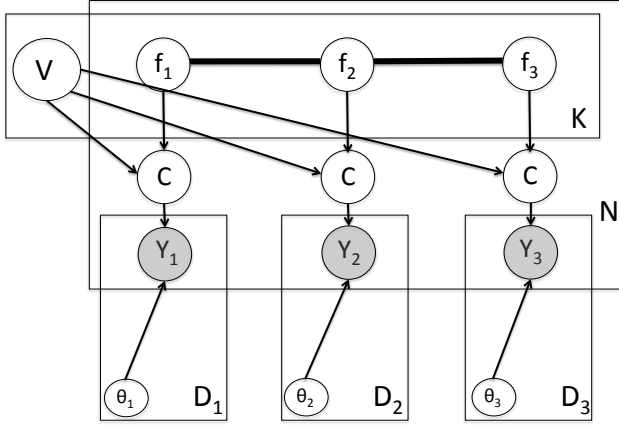


Figure 1. Graphical model for both the Multitask Clustering Model (MCM) and Evolving Community Structure (ECS) model. c are a deterministic function of f and v but we show both for clarity. The thick line linking the f 's denotes dependence through the Gaussian process.

4. Multitask clustering

We are interested in the situation where we have a collection of N objects, for each of which we have measurements from T different data sources. We associate each data task τ with a covariate location $t_\tau \in \mathcal{T}$, although this might be arbitrary. Our model will assume that for each data source the objects are grouped into clusters forming a partition. We allow the clustering to be different for each data source, but model dependency between these partitions using the DPVP. Denote the data for object n in data source τ as $y_n^\tau \in \mathcal{Y}^\tau$, where we allow the observed space \mathcal{Y}^τ to be different for each data source. The Multitask Clustering Model (MCM) is then

$$\begin{aligned} c_n^\tau &\sim DPVP(\alpha, \mathbf{t}, \Sigma) \\ \theta_k^\tau &\sim G^\tau \\ y_n^\tau | c_n^\tau, \theta_k^\tau &\sim F^\tau(\theta_{c_n^\tau}^\tau) \end{aligned} \quad (6)$$

where θ_k^τ are cluster parameters, G^τ are priors on the cluster parameters and F^τ are data likelihoods. In the following we assume all the data sources are continuous, i.e. $\mathcal{Y}^\tau = \mathbb{R}^{D^\tau}$, and can therefore be represented as a $N \times D^\tau$ matrix $\mathbf{Y}^\tau \in \mathbb{R}^{N \times D^\tau}$. We allow each data source to have a different observed dimensional D^τ . We use a diagonal (independent per dimension) Gaussian likelihood and its conjugate normal-gamma prior on the cluster parameters:

$$\begin{aligned} (\mu_{kd}^\tau, \lambda_{kd}^\tau) &\sim \mathcal{N}(\mu_{kd}^\tau | \mu_o, \frac{1}{\kappa_o \lambda_{kd}^\tau}) \mathcal{Ga}(\lambda_{kd}^\tau | \alpha_o, \frac{1}{\beta_o}) \\ y_{nd}^\tau | c_n^\tau, \mu, \lambda &\sim \mathcal{N}(\mu_{c_n^\tau d}^\tau, 1/\lambda_{c_n^\tau d}^\tau) \end{aligned} \quad (7)$$

where we set $\kappa_o = 0.1, \mu_o = 0, \beta_o = 0.1, \alpha_o = 0.1$. By choosing the conjugate prior we are able to integrate out the cluster parameters during inference. The choice of a diagonal covariance allows scaling to larger datasets. The generalisation to other data types would be straightforward using the appropriate conjugate prior.

5. A model for time evolving community structure

Most models of network data assume a static network, whereas real world network typically evolve over time. We use the DPVP to build a model which discovers clusters in network data, but allows cluster assignments to change over time. We refer to this model as ECS for ‘‘Evolving Community Structure’’. In this case our data \mathbf{Y}^τ at each location τ is a binary $N \times N$ matrix representing the presence or absence of links between objects. The assignment of the objects to groups at each covariate index t determines the probability of links in an analogous fashion to the Infinite Relationship Model (IRM, Kemp et al., 2006).

$$c_n^\tau \sim DPVP(\alpha, \mathbf{t}, \Sigma) \quad (8)$$

$$\theta_{kk'}^\tau \sim \text{Beta}(\beta, \beta) \quad (9)$$

$$y_{nn'}^\tau | c^\tau, \theta^\tau \sim \text{Bernoulli}(\theta_{c_n^\tau c_{n'}^\tau}^\tau) \quad (10)$$

where $y_{nn'}^\tau$ denotes the presence of a link between objects n and n' at time τ , and we set $\beta = 0.1$ to encourage values of θ close to 0 or 1. While we assume the link probabilities θ are independent at every time point τ a possible extension would be to introduce dependence between these values at adjacent time points.

6. Choice of kernel

Since the DPVP is constructed using Gaussian processes there is great flexibility in the choice of covariance function (kernel). Here we describe only the options used in the experiments in Section 8. We choose to fix the diagonal $\Sigma(\tau, \tau) = 1$ in all cases since the DPVP is invariant to scaling of the covariance matrix: by construction changing the prior marginal variance of the GP functions at any location does not effect the resulting distribution over partitions.

Squared exponential. In the situation where there is a known covariate value, such as time or spatial location, associated with each data source or network, the canonical covariance function is the squared exponential kernel

$$\Sigma(t, t') = \exp\left(-\frac{(t - t')^2}{2l^2}\right)$$

where $l > 0$ is the lengthscale which controls the smoothness of the GPs. We put an exponential prior on l .

Similarity kernel. In the multitask clustering setting we may often have no little prior knowledge about which data sources are likely to have similar clustering structure to others and would like to learn this from the data itself. In this case we fix the diagonal terms equal to 1 and put a Uniform $[-1, 1]$ prior on the off-diagonal terms. We ensure PSD matrices simply by rejecting any non-PSD matrices during the slice sampling.

Tree structured covariance. In other cases we may have a tree structured dependency between the data sources: the data sources are the leaves of a tree which represents a know relationship. Each branch in the tree represents a Gaussian factor: the child node is normally distributed with mean equal to its parent and variance equal to the branch length. The total height of the tree is 1 so that the diagonal of the resulting covariance matrix is 1. We put a uniform prior on the branch lengths (under the constraint that no branch lengths can be negative). A concrete example is given in Section 8.3.

7. MCMC Inference

In the following, we present a method for inferring the latent variables and parameters of MCM and ECS model: the matrix of Gaussian process function values, \mathbf{F} , the weight vector \mathbf{v} and the parameters of the covariance matrix Σ . Exact inference is intractable, so we develop a Markov Chain Monte Carlo (MCMC) procedure to sample from the posterior distribution. Each iteration is $O(DNKT + T^3)$ which for small T is the same as EM or sampling for a DPM. The cluster assignments c are not represented since these are a deterministic function of \mathbf{F} and \mathbf{v} (since we ensure that $\Sigma_{\tau\tau'} = 1$ in all cases the assignments do not depend on Σ). For both models we are able to integrate out the cluster parameters, θ (the link probabilities for ECS), due to conjugacy:

$$p(\mathbf{Y}^\tau | \mathbf{F}^\tau, \mathbf{v}) = \int p(\mathbf{Y}^\tau | \mathbf{F}^\tau, \mathbf{v}, \boldsymbol{\theta}^\tau) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad (11)$$

The sampler iterates as follows:

Sampling the GP function values, \mathbf{F} . The conditional posterior over the matrix \mathbf{F} is given by

$$p(\mathbf{F} | \mathbf{Y}, \mathbf{v}, \Sigma) \propto \left(\prod_{\tau=1}^T p(\mathbf{Y}^\tau | \mathbf{F}^\tau, \mathbf{v}) \right) \prod_{n=1}^N \prod_{k=1}^{K-1} N(\mathbf{f}_{nk} | 0, \Sigma)$$

Since we have a GP prior over \mathbf{F} we use elliptical slice sampling (ESS, Murray et al., 2010), which is specifically designed to sample from posteriors with strongly correlated Gaussian priors. We find that jointly sampling the $K - 1$ GPs associated with a datapoint gives better mixing than attempting to jointly sample all $N(K - 1)$ GPs jointly (see Section 8.1). The covariance matrix for the $K - 1$ GPs for a single datapoint is block diagonal where each block is Σ . Naive computation of the Cholesky would be $O(K^3 T^3)$, but utilising the block diagonal structure it is only $O(T^3)$, and T is typically relatively small in the applications we envisage.

Sampling the weight vector, \mathbf{v} . The sampler successively samples each of the $K - 1$ weights v_k . Since we do not have conjugacy (due to the complex form of likelihood function), we cannot sample directly from the posterior $v_k | \mathbf{Y}, \mathbf{F}, \mathbf{v}_{-k}$ (where \mathbf{v}_{-k} denotes the values of \mathbf{v} excluding v_k). To overcome this problem, we use slice sampling (Neal, 2003b) with the reparameterisation $g(v) = \log[v/(1 - v)]$ so that $g \in \mathbb{R}$.

Sampling parameters of the kernel. Learning the parameters of the kernel is of interest because it tells us how similar the partitions appear to be across covariate space. Since we ensure $\Sigma_{\tau\tau'} = 1$ the likelihood for Σ is just $\prod_{n=1}^N \prod_{k=1}^{K-1} N(\mathbf{f}_{nk} | 0, \Sigma)$. In fact, since values of \mathbf{f}_{nk} for $k > \max c$ do not effect the partitioning, we can trivially integrate these out so the the product over k in the likelihood need only go up to $\max c$. This both saves computation and improves mixing by avoiding conditioning on irrelevant information. For all three kernels described in Section 6 we use slice sampling to learn the kernel parameters: the lengthscale for the squared exponential kernel, the correlation coefficients for the similarity kernel, and the branch lengths for the tree structured kernel.

Sampling the concentration parameter, α . We also use slice sampling to infer the hyperparameter α using a Gamma prior $\alpha \sim \mathcal{G}(1, 1)$. The likelihood is $\prod_{k=1}^{K-1} \text{Beta}(v_k | 1, \alpha)$.

Initialisation. The DPVP can suffer from the label switching problem: although the partitioning at two locations might be similar, they may look quite different according to the model if the labels are permuted. To alleviate this problem we initialised both MCM and ECS using the equivalent DPM model where the same clustering is shared across all locations.

8. Multitask clustering results

Using synthetic data we first demonstrate some encouraging characteristics of our inference method. We then apply MCM to two real world biological datasets. Unless otherwise stated we use 1000 MCMC iterations, discarding the first 500 as burnin.

8.1. Synthetic data

We use experiments on synthetic data to demonstrate three things. Firstly, on a single location dataset with $N = 30$ objects in three equal sized, well separated clusters with means $-31, 0, +31$ and equal covariances I in $D = 5$ dimensions, we show that our Elliptical Slice Sampling (ESS) based inference is competitive with standard Gibbs sampling for DPMs (for $T = 1$ MCM is exactly equivalent to a DPM). In Figure 2 we see that while ESS does sometimes get stuck at in a local mode (two of ten repeats), it typically finds areas of higher marginal likelihood than the standard Gibbs sampler.

Secondly, on a dataset with $T = 3$ data sources, two of which have the same clustering structure as the first example, and one of which has a single cluster with mean 0 and variance I , we show that sequentially jointly sampling the $K - 1$ GPs associated with each datapoint mixes better than attempting to sampling all $N(K - 1)$ GPs and \mathbf{v} jointly (Figure 3). The latter approach attempts to make very large global moves in the MCMC state space so many likelihood evaluations are required before a new point is accepted by ESS.

Thirdly, on the same dataset we show we can learn a sensible similarity kernel. We use the similarity kernel defined in Section 6. We run MCMC for 200 iterations and, finding the chain appears to have burnt in after 100 iterations, calculate 95% credible intervals for Σ using the final 100 samples. The correlation coefficient between the two data sources whose true clusterings are identical has credible interval is $[0.31, 0.65]$, whereas between the distinct data source and these two the credible intervals are $[-0.39, 0.27]$ and $[-0.37, 0.23]$.

8.2. Cancer cell line encyclopedia

The Cancer cell line encyclopedia (CCLE, Barretina et al., 2012) is a recently published resource to aid the understanding of why certain cancer subtypes are resistant to particular drugs. $N = 432$ cancer cell lines were grown in the presence of 24 different therapeutic compounds at nine different concentrations, and their growth measured after 10 days. These growth curves are summarised in terms of “active area”: the

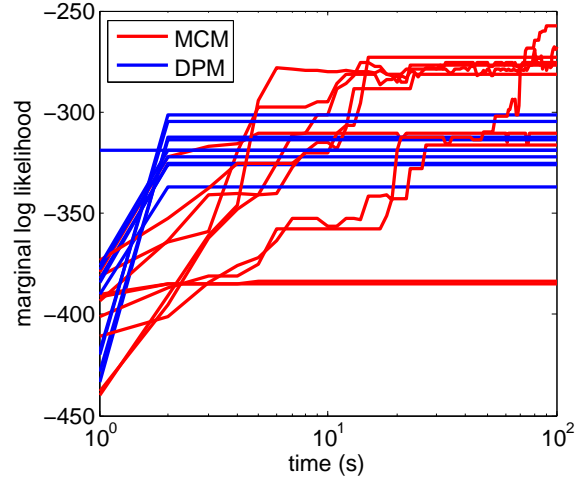


Figure 2. Comparison of our ESS based inference vs standard Gibbs sampling for the DPM when there is only $T = 1$ location.

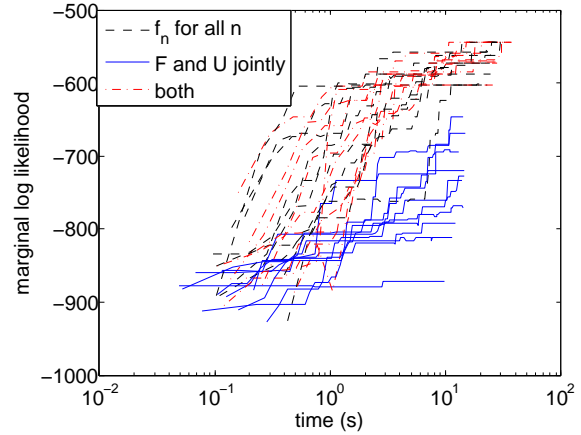


Figure 3. Comparison of sampling methods for \mathbf{F} . Dashed black: sequentially sampling the $K - 1$ GPs associated with each object n sequential, followed by \mathbf{v} . Solid blue: sampling all of \mathbf{F} and \mathbf{U} jointly. Dashed red: alternating between these two.

total inhibition of growth summed over all concentrations. Alongside these sensitivity measurements various molecular characteristics of the cell lines are measured, including gene expression (GE), copy number variation (CNV, the number of times a gene is duplicated in the cancer genome) and oncogene mutations (mutations such as insertions, deletions or single nucleotide polymorphisms, SNPs, in genes known to be involved in cancer). We consider two tasks: firstly, predicting drug sensitivity, since this is clinically relevant (being able to predict sensitivity could help determine what drug is most appropriate for a particular patient) and secondly, learning how the clustering of cancer cell varies across the four data sources: GE, CNV, oncogene mutations (ONCO) and sensitivity (SENS).

To assess the predictive performance of MCM on drug sensitivity, we randomly choose 20 different sets of 10% of the sensitivity measurements to hold out, and attempt to impute these values. We compare to two extremes: using a DP mixture (DPM) model independently on each data source (in this case we need only run the algorithm for the sensitivity data source since this is what we are interesting in imputing), and a DPM where the clustering is common to all data sources. These are extremes of MCM, representing minimum or maximum transfer learning respectively. The results are shown in Table 1. We see that the independent clustering performs very poorly, the shared clustering does reasonably well, but MCM performs best since it is able to learn how much information to transfer from the other data sources to the drug sensitivity clustering task. The learnt correlation matrix is shown in Figure 4: of particular interest are the correlation coefficients between sensitivity and the other data sources. We see there is a positive correlation to CNV, whereas the correlation is small (and in fact slightly negative) to gene expression and oncogene mutations, suggesting that copy number variation is the most indicative characteristic of which drugs a cancer will be sensitive/resistant to.

Dataset	Independent	Shared	MCM/ECS
CCLE	-3.221 ± 0.552	-1.109 ± 0.069	-0.902 ± 0.097
van de Bunt	-0.530 ± 0.025	-0.502 ± 0.022	-0.095 ± 0.017
HapMap	-1.357 ± 0.055	-1.134 ± 0.013	-1.277 ± 0.016

Table 1. Predictive performance results for both MCM and ECS on real world datasets. Values are log predictive likelihood per heldout data entry.

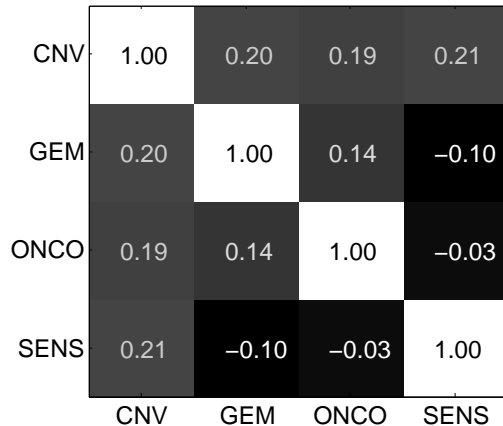


Figure 4. Correlation matrix learnt for different data sources in the CCLE dataset.

8.3. HapMap gene expression data

The HapMap project¹ is primarily an attempt to measure genetic variation between 1301 individuals from different human populations, but gene expression data is also available in 618 individuals (Montgomery et al., 2010). We consider the task of discovering regulatory modules of genes from this gene expression data, but rather than simply learning a global clustering we will use the MCM to learn population specific clusterings of the genes. From around 20,000 genes we filter down to the 1000 most variable ones. The known tree over the different populations is shown in Figure 5. We use this tree to define the covariance matrix as described in Section 6.

We again assess predictive performance on 10 heldout sets consisting of 10% of data entries. In this case we find that although MCM performs better than independent clustering in each population, using a shared clustering of genes across all populations performs best. This suggests the biological conclusion that gene regulatory modules do not vary between diverse human populations.

9. Network modelling results

We experimentally evaluate the ECS model on both synthetic and real-world data.

9.1. Synthetic data

We explore the ability of ECS to recover the partitions in synthetic time series network data. We hand-constructed a set of six square binary matrices which

¹<http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html>

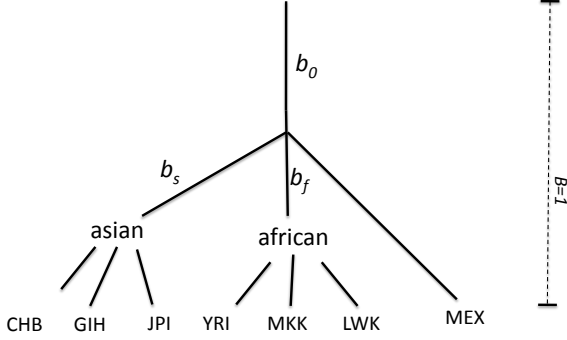


Figure 5. Tree structure of human populations in HapMap.

encode the friendship links among $N = 30$ people evolving through time, as shown in Figure 6. People form groups (clusters) which determine the links and non-links between them. As time passes, the partitioning of people changes; new friendship links are created while others break. The closer in time two snapshots are, the more similar we expect the related partitions will be. We ran ECS for 200 MCMC iterations and the sample with highest marginal likelihood is shown in Figure 7. We see that the solution provided by our model proposes two clusters at $t = 0$, which shrink as a third cluster is generated between them. The partition found at $t = 3$ is suboptimal: with more iterations we might hope the white and yellow clusters used here would be replaced by the red cluster used at latter time points. However, multiple hypotheses are of course capable of explaining the same data.

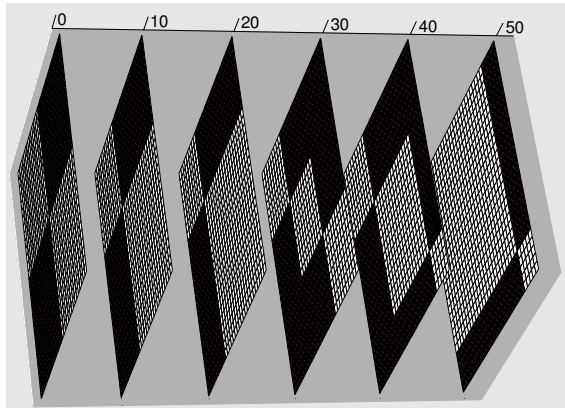


Figure 6. Synthetic network data. Each matrix represents (*non*-)links among pair of objects. White corresponds to one (link) and black to zero (*non*-link).

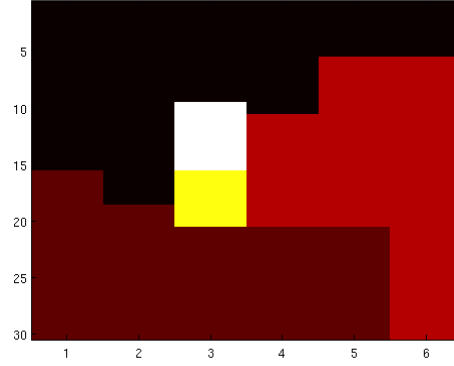


Figure 7. Learnt clustering for synthetic network data. Colours denote assignment.

9.2. van de Bunt’s students

In [Van De Bunt et al. \(1999\)](#) 32 university freshman were surveyed at seven time points about who in their class they considered as friends. The first four time points were two weeks apart and the last four being three weeks apart. We binarise the original 0 – 5 scale, taking ‘Best friendship’, ‘Friendship’ and ‘Friendly relationship’ as 1, ‘Neutral relationship’, ‘Unknown person’ and ‘Troubled relationship’ as 0, and ‘non-response’ as missing. We also symmetrise the matrix by assuming friendship if either individual reported it. We test the predictive performance of ECS using the squared exponential kernel on this dataset by holding out 10% of all links across all time points in 10 different training/test splits. We compare to independent IRMs at each time point and a model with shared clustering but independent link probabilities across all time points. The results in Table 1 show that ECS significantly outperforms both these extremes in terms of heldout predictive performance. The average length-scale learnt was 1.43 weeks, showing that while there was similarity in community structure between proximal time points there were also significant changes to be taken into account over the time course of the full dataset.

10. Conclusion

Given the central role of clustering in unsupervised learning we expect the dependent partitioned-valued process we introduce here to have many potential applications. We have investigated two models derived from the DPVP: a multitask clustering model, which is, to the best of our knowledge, the first such model derived under a fully probabilistic framework, and a time series network model that is appropriate for

the many real world networks that constantly evolve through time.

Various directions for future work are open. Firstly, improved inference is of interest. While we used MCMC inference, variational methods would be a natural fit for either model: expectation propagation (Minka, 2001) is known to be particularly effective for Gaussian process classification (Nickisch & Rasmussen, 2008), which is a subcomponent of DPVP, and variational Bayes (Attias, 2000; Ghahramani & Beal, 2001) is commonly used for mixture modelling. Secondly, it is possible to have covariates associated with each object n . Making the f also a function of these covariates would give a model related to the distance dependent CRP (Blei & Frazier, 2011), and smart computation of the Cholesky would be only $O(T^3 + N^3)$ rather than the naive $O(T^3 N^3)$. Thirdly, other applications suggest themselves: modelling spatially varying ecological networks or the difference between regulatory modules across different human tissue types. Finally it would be of considerable interest to derive a dependent partition-valued process that, like the FCP, does not explicitly label clusters, and therefore does not suffer from the label switching problem.

References

- Aldous, D J. Exchangeability and related topics. In *Ecole d'Ete de Probabilités de Saint-Flour*, volume XIII, pp. 1–198. Springer, 1983.
- Attias, H. A variational Bayesian framework for graphical models. In *Advances in Neural Inf. Proc. Systems (NIPS) 12*, 2000.
- Barretina, Jordi, Caponigro, Giordano, Stransky, Nicolas, Venkatesan, Kavitha, Margolin, Adam A, Kim, Sungjoon, Wilson, Christopher J, Lehár, Joseph, Kryukov, Gregory V, Sonkin, Dmitriy, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 2012.
- Bertoin, Jean. *Random fragmentation and coagulation processes*, volume 102 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2006.
- Blei, David M and Frazier, Peter I. Distance dependent Chinese restaurant processes. *The Journal of Machine Learning Research*, 2011.
- Duan, A., Guindani, Michele, and Gelfand, Alan E. Generalized spatial dirichlet process models. In *Duke University*, pp. 05–23, 2005.
- Ferguson, T S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1973.
- Ghahramani, Zoubin and Beal, Matthew J. Propagation algorithms for variational Bayesian learning. In *Advances in Neural Inf. Proc. Systems (NIPS) 13*, 2001.
- Ishwaran, Hemant and James, Lancelot F. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173, 2001a.
- Ishwaran, Hemant and James, Lancelot F. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 2001b.
- Kemp, C, Tenenbaum, J B, Griffiths, T L, Yamada, T, and Ueda, N. Learning systems of concepts with an infinite relational model. In *Proc. of the National Conf. on Artificial Intelligence*, 2006.
- Kingman, J F C. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.
- Kirk, Paul D. W., Griffin, Jim E., Savage, Richard S., Ghahramani, Zoubin, and Wild, David L. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, 2012.
- MacEachern, S. N. Dependent nonparametric processes. In *Proc. of the Section on Bayesian Statistical Science*, pp. 50–55. American Statistical Association, 1999.
- MacEachern, S. N. Dependent dirichlet processes. Technical report, Ohio State University, 2000.
- MacEachern, Steven N. Dependent nonparametric processes. In *ASA Proc. of the Section on Bayesian Statistical Science*. American Statistical Association, 1999.
- Minka, T P. Expectation propagation for approximate Bayesian inference. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 17, 2001.
- Montgomery, Stephen B, Sammeth, Micha, Gutierrez-Arcelus, Maria, Lach, Radoslaw P, Ingle, Catherine, Nisbett, James, Guigo, Roderic, and Dermitzakis, Emmanouil T. Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, 2010.
- Murray, Iain, Adams, Ryan Prescott, and MacKay, David J.C. Elliptical slice sampling. *JMLR: W&CP*, 9:541–548, 2010.

- Neal, R. M. Density modeling and clustering using Dirichlet diffusion trees. *Bayesian Statistics*, 7:619–629, 2003a.
- Neal, Radford M. Slice Sampling. *The Annals of Statistics*, 2003b.
- Nickisch, Hannes and Rasmussen, Carl Edward. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9: 2035–2078, October 2008.
- Rasmussen, Carl Edward and Williams, Christopher K I. *Gaussian processes for machine learning*. MIT Press, 2006.
- Sethuraman, Jayaram. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.
- Teh, Y. W., Blundell, C., and Elliott, L. T. Modelling genetic variations with fragmentation-coagulation processes. In *Advances In Neural Inf. Proc. Systems*, 2011.
- Van De Bunt, Gerhard G, Van Duijn, Marijtje AJ, and Snijders, Tom AB. Friendship networks through time: An actor-oriented dynamic statistical network model. *Computational & Mathematical Organization Theory*, 1999.
- Williamson, Sinead, Orbanz, Peter, and Ghahramani, Zoubin. Dependent Indian buffet processes. In *Proc. International Conference on Artificial Intelligence and Statistics*, 2010.